Post-print of: A. Kwasniewska, J. Ruminski, M. Szankin and K. Czuszynski, "Pose-Invariant Face Detection by Replacing Deep Neurons with Capsules for Thermal Imagery in Telemedicine," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, 2018, pp. 778-781, doi: 10.1109/EMBC.2018.8512381.

Pose-Invariant Face Detection by Replacing Deep Neurons with Capsules for Thermal Imagery in Telemedicine

A. Kwasniewska, *Student Member IEEE, EMBS*, J. Ruminski, *Member IEEE,*
M. Szankin, *Member IEEE and* K. Czuszynski

***Abstract*— The aim of this work was to examine the potential of thermal imaging as a cost-effective tool for convenient, non-intrusive remote monitoring of elderly people in different possible head orientations, without imposing specific behavior on users, e.g. looking toward the camera. Illumination and pose invariant head tracking is important for many medical applications as it can provide information, e.g. about vital signs, sensory experiences, injuries, wellbeing. In the performed experiments, we investigated the influence of different modifications of images (rotation, displacement of facial features, and displacement of facial quarters) on the prediction accuracy. Specifically, two models were tested on the set of collected low-resolution thermal images: Inception V3 Convolutional Neural Network (CNN) and Hinton's Capsule Network. The preliminary results confirm that the prediction ability of the model based on capsules can deal with different head orientations much better than CNN (for the 45º head rotation Capsule Network achieved ~100% accuracy while CNN only 9.5%).**

I. Introduction

In view of rapidly aging societies all over the world [1] and the incremental cost of health care as a proportion of the change in GNP [2], more and more home-based health care solutions have been considered as a new frontier in medical practice, e.g. monitoring of physiological parameters with computer vision module [3], self-diagnostics with smart glasses [4] or a web-based telemedicine system [5]. Numerous applications are focused on investigating changes that appear within a face, as it is an overly sensitive region that exposes a lot of information about vital signs [6], sensory experiences [7], injuries [8][9] or wellbeing [10]. Image processing help to eliminate the usage of additional sensors. Typically, breathing is measured with the means of chest-belts [11], but it has been proved that thermal imaging can support analysis of breathing patterns as well by investigating temperature changes in the nostril area for e.g. emotions detection [12]. In this work, we would like to extend our research on remote respiratory rate evaluation presented in [6], by providing accurate algorithm for automatic face tracking that will work regardless of camera angle or body orientation, as our previous work was limited to manual selection of region of interest (ROI).

Our study can also enable our remote medical diagnostic solutions. For example, apart from breathing rate, an important vital sign that can be calculated by processing images of the facial region is a pulse rate [13]. Face can be also analyzed for pain assessment [14] or paralysis grading [15]. Locating face coordinates is an important prerequisite to make these telemedicine systems fully automatic. Most of the existing solutions, though, are based on the assumption that a patient is keeping his head straight. Our daily routine requires to support other scenarios that do not impose specific behaviors on users, e.g. tilting head sideways, lying down, also considering poor lightning conditions.

Recent methods for face detection were mainly based on Convolutional Neural Networks (CNN), as they significantly outperform previous approaches mostly based on hand-crafted features [16]. Li H. et al. [17] aimed at differentiating faces from the background by using a cascade architecture built on CNNs that evaluate low resolution images first to quickly eliminate non-facial regions. The rotational invariance was discussed in Deep Dense Face Detector [18]. The proposed detector uses CNN network to extract features and classify a face in a wide range of orientations. Deep models can also be utilized for facial landmark alignments to estimate the position of a face and address its variations [19]. Although the implementation of more advanced systems that make use of a computer vision became feasible due to advantages in deep learning, pose and poor lighting is still a standing problem for CNN [20] (see Fig. 1). Obviously, there are various pose estimation methods [21] but it increases computational complexity and usually additional sensors are required for high accuracy, e.g. RGB-D Sensor [22].

a) Rotation 45° b) Rotation 90° c) Rotation 180°

Figure 1. Row 1st and 2nd: examples of cases that should be supported in telemedicine systems; row 3rd: examples of IR images corresponding to above scenarios chosen from the test set; accuracy for the proposed approach a) 100%; b) 78%; c) 99%; CNN a) 9%; b) 20%; c) 84%

This work has been partially supported by NCBiR, FWF, SNSF, ANR and FNR in the framework of the ERA-NET CHIST-ERA II, project eGLASSES – The interactive eyeglasses for mobile, perceptual computing and by Statutory Funds of Electronics, Telecommunications and Informatics Faculty, Gdansk University of Technology.

A. Kwasniewska (alicja.kwasniewska@pg.edu.pl), J. Ruminski (jacek.ruminski@pg.edu.pl) K. Czuszynski (krzczusz@student.pg.gda.pl) are with Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Deptartment of Biomedical Engineering, Gdansk, Poland

A. Kwasniewska (alicja.kwasniewska@intel.com), M. Szankin (maciej.szankin@intel.com) are with Intel Corp., San Diego, CA, USA

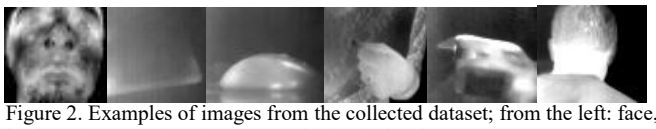

Figure 2. Examples of images from the collected dataset; from the left: face, keyboard, mouse, hand, projector, back of a head

In our solution, we want to preserve a minimal cost of the system and make it as imperceptible as possible to provide a non-intrusive way for remote monitoring of people. Therefore, we propose to use a low-cost thermal camera module for thermal face detection and tracking to capture in a compact way the variability of vital signs (respiratory rate [6].) visible in the facial area. Thermal imaging proposed for this use case allows for eliminating lighting and privacy concerns. Because of the small camera size, it could be potentially embedded in a smart home infrastructure or wearable devices, e.g. eGlasses platform [23]. To be able to run our solution on resource-constrained devices, we want to make it as simple as possible and do not apply other image processing techniques for e.g. pose estimation, what will increase the requirement for memory footprint and computational resources.

Our second contribution lays in the use of Capsule Networks [24] to reduce the influence of pose changes on the system accuracy. As discussed in [25] the lack of rotational invariance in CNN can cause the model to produce incorrect predictions. In our previous studies, we showed that detection can be performed by removing the final pooling operations and modifying the CNN network architecture during the inference [26]. However, in our previous work we assumed strictly defined conditions of data acquisition, where person is looking towards the camera. To deal with real life scenarios, the designed algorithm should be not sensitive to face orientations, what we want to address in this paper. The concept of Capsules proposed by Hinton et al. [24] uses dynamic routing to make the model rotationally invariant and spatially aware. This architecture has already been successful on RGB data (e.g. MNIST, CIFAR10 [25]), yet it has not been applied to thermal IR images.

Additionally, we compare the reliability of the proposed approach with the core architecture used for classification – Inception v3 [27]. Specifically, we test if a face can be accurately classified using CNN and Capsule Network from low-resolution thermal images in uncontrolled face detection problems, where face rotations can lead to significant changes in visual appearance and therefore degrade the robustness of the classificator.

The rest of the paper is organized as follows: Section II demonstrates technical details of the proposed method applied on thermal images. Next, we summarize the experimental results in Section III. Finally, we discuss and compare of our new approach against existing solutions in Section IV and conclude the paper in Section V.

II. METHODOLOGY

In this section, we present details of the algorithm and training process of the proposed thermal face classificator based on the Capsule Network architecture [24]. Additionally, we compare our approach with another network architecture (Inception [27]) that has been a core model used for the classification over past few years.

In our experiment, we trained two models: Inception v3 and Capsule Network on images with the face placed straight to ease the data collection process. Various modifications (face rotation and facial features displacement) were introduced later to test the accuracy of both solutions in different possible scenarios of telemedicine examinations. For data collection, the FLIR® Lepton thermal camera module characterized by 14-bits dynamic range, size <1cm2 and spatial resolution of 80x60 was used. Sequences of frames were captured for 26 healthy volunteers (age: 26.8±8.1) in a laboratory room at an ambient temperature 23–27 °C during 60s period (sampling frequency fs=12Hz) at a distance ~0.4-1m. From the acquired recordings, we extracted 3256 images of a face. Additionally, we collected thermal images of objects present in the laboratory room and other body parts, creating another 5 categories (mouse – 2855 images, projector - 2968, keyboard - 3086, back of a head - 3083, hand – 3083; see Fig. 2). Analyzing the collected set of faces images, we realized that a contrast of facial features was much lower than in the visible light images what made the interpretability of thermal images difficult. Therefore, we applied a pre-processing technique based on fitting the Gaussian distributions to histogram data and scaling pixels' values representing the face area to the range of 0-255, as described in [26]. As a result, usable data previously represented by close values gained a higher contrast and it was easier for models to learn features, as their architecture is based on extracting high frequency components (edges, corners, etc.). The examples were divided into train, validation and tests sets (0.8:0.1:0.1).

We started by training the Capsule Network with the collected dataset. This architecture was described in detail in [24]. Generally, it is based on the idea of dividing each layer in a model into 'capsules' that are small groups of neurons, where the input and output of the capsule is represented by a vector, not a scalar as in traditional neural networks. Initially, the output of the capsule i in layer n (Cap_i^n) is sent to all capsules in layer $n+1$ ($Cap_{1\dots m}^{n+1}$), where m represents number of capsules. For each of $Cap_{1\dots m}^{n+1}$ the prediction vector ($u_{1\dots m}$) is calculated by multiplying the output of Cap_i^n by the weight matrix. The u that produces the largest scalar product with the output of $Cap_{1\dots m}^{n+1}$ is chosen (u_z) and the top-down feedback is applied to increase the coupling coefficient (c_z) for this capsule Cap_z^{n+1}. It has been already demonstrated [24] that this mechanism known as iterative routing-by-agreement is more effective than max-pooling used in CNN, which takes into account only the most active feature. The probability that the entity is present in the input is represented by a length of output vector of capsule i (v_i^{n+1}) defined as:

$$v_i^{n+1} = \frac{\|s_i^{n+1}\|^2}{1+\|s_i^{n+1}\|^2} \frac{s_i^{n+1}}{\|s_i^{n+1}\|} \quad (1)$$

where s_i^{n+1} is the total input to capsule i in layer $n+1$, defined as a sum over all $u_{1\dots m}$ weighted by $c_{1\dots m}$. As suggested in [24] m^+=0.9, m^-=0.1 and regularization parameter λ=0.5 was used for calculating a loss for class k:

$$L_k = T_k \max\left(0, m^+ - \|v_k^t\|\right)^2 + \lambda(1-T_k) \max\left(0, \|v_k^t\| - m^-\right)^2 \quad (2)$$

where T_k=1 if class k is present in an image and v_k^t is the output vector of capsule for class k in top layer t. The

network was trained using 1, 3, 4 and 7 routing iterations in each case for 50 epochs.

With the same dataset, we trained Inception v3 network, which architecture is explained in details in [27] using transfer learning technique as described in [26] for 20000 steps with learning rate set to 0.01. To examine and compare the tolerance of both networks for possible uncontrolled face detection problems, we introduced 5 modifications (Fig. 3) using 159 images for each case: M1: random displacement of facial features; M2: random displacement of image quarters; M3: rotation 90°; M4: rotation 180°; M5: rotation 45°. The flow of the experiment procedure is presented in Fig. 4.

III. Results

After training, the final accuracy of models was calculated on the test set (Table I). Then, 159 random images from the face class were selected and modified. Table II presents the percentage of images classified as a face in a given case. For each image classified to the face class in each scenario the probability that this image belongs to the 'face' category was also computed. Average of all calculated probabilities (± standard deviation) is presented in the bottom half of cells in the Table II. In each row of Table II we highlighted cells that correspond to the best results (the darker the better). Baseline (B) represents results for selected 159 images of a face without modifications. Modification 1, 2 are 'negative' cases: faces were distorted, so the number of samples classified as 'face' should be the smallest. Modification 3, 4, 5 are 'positive' cases: introduced rotations should not impact the number of proper predictions (number of samples classified as a face should be the highest).

TABLE I. Accuracy of the test set [%]

Inception	Capsule Network			
	iter. rout. 1	*iter. rout. 3*	*iter. rout. 4*	*iter. rout. 7*
98.91	99.92	99.85	99.88	99.66

TABLE II. Percentage of images classified as 'face' [%] and the avg. probability [%] of belonging to this class (avg ± stdev)

	Inception	Capsule Network			
		iter. rout. 1	*iter. rout. 3*	*iter. rout. 4*	*iter. rout. 7*
B	94.68 96.04±7.79	100 91.20±1.35	100 79.4±3.36	100 65.47±4.73	100 58.93±5.60
M1	87.34 92.87±7.66	99.28 89.94±2.63	99.36 74.5±5.54	99.36 62.00±5.87	100 54.44±6.8
M2	48.10 77.66±19.20	54.43 79.1±8.23	46.84 59.00±11.53	48.03 45.43±11.14	56.32 39.32±10.58
M3	20.25 69.37±14.78	78.48 76.64±7.89	67.72 57.27±9.87	70.88 42.92±12.08	82.91 38.55±9.66
M4	83.54 85.79±14.32	99.36 86.80±4.10	96.83 67.92±7.90	96.83 54.64±9.29	98.73 47.97±9.86
M5	9.49 63.25±9.41	100 83.42±3.35	99.36 56.67±6.82	91.14 43.08±7.44	99.36 45.71±6.76

IV. Discussion

In our experiments, at first, we measured the performance achieved by the CNN and the Capsule Network in all tested configurations on a test set. Results show that both models achieve high classification accuracy for images where a person is looking towards the camera. It was also confirmed that the more iterations in the dynamic routing mechanism, the more probable that the network will overfit to the training set. We observed that with more iterations, the accuracy was decreasing.

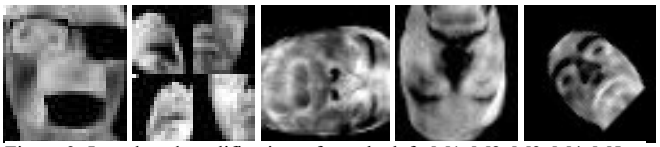

Figure 3. Introduced modifications; from the left: M1, M2, M3, M4, M5

The goal of our study was to determine if a face can be accurately classified using CNN and Capsule Network from low-resolution thermal images in uncontrolled daily scenarios, where face rotations can lead to significant changes in visual appearance and therefore degrade the robustness of the classificator. As presented in Table II (M3, M4, and M5) the Capsule Network can deal with different face orientations much better than CNN. For all rotations, architecture based on capsules properly classified a face in much more images than deep neural network. For rotation 90° the best results were achieved for the network trained using 7 iterations; for rotation 180° and 45° 1 iteration. CNN was able to distinguish a face in less than 10% of images rotated by 45° and ~20% of images rotated by 90°. Considering rotation invariance problems, the CNN performance could be potentially improved by training the network with augmented data using rotated images, what we want to explore in the future. Yet, taking into account the target platform, which include smart home devices, the setup time (e.g. training the model to recognize specific person) should be minimal, while bigger dataset would significantly increase the training time.

In case, when a face was distorted (M1, M2) both networks produced average results. In modification based on facial features displacement (M1), images were classified as a face in more than 87% examples for CNN and 99% for Capsule Network. However, we should consider the low resolution of the collected images, what causes the features to be significantly blurred. In most cases, it was impossible to distinguish specific part of a face even by a human. This could have led to improper predictions. On the other hand, for modification based on displacement of image quarters, Capsule Network was slightly better than the CNN.

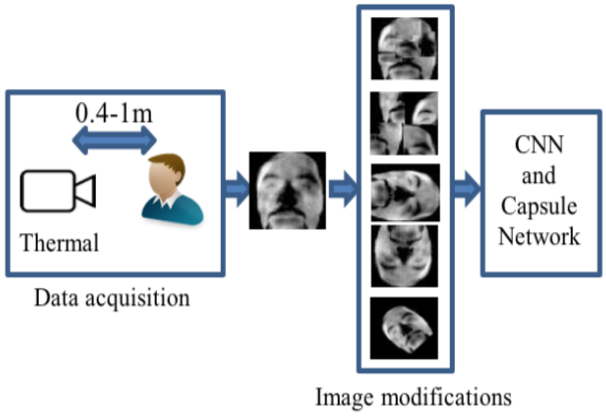

Figure 4. The flow of the experiment procedure

Over the past few years, CNNs have revolutionized many computer problems, increasing accuracy of image classification models beyond human capabilities. In our experiments, we compared the popular deep neural model with the Capsule Network. As already mentioned, reliable face tracking in various real-life scenarios enables many remote healthcare systems that can improve the quality of elderly people life and, more importantly, detect emergency situations, e.g. remote prediction of cardiovascular disease by analysis of thermal images of a face [8].

V. Conclusion

This work was focused on improving remote respiratory rate estimation study [6] by providing automatic, rotation invariant algorithm for face detection. The proposed method was analyzed on a variety of thermal face images modified to simulate possible real-life applications of telemedicine systems. According to preliminary results, the proposed capsule-based thermal face classification is able to handle the tested scenarios and accurately detect a face regardless of its orientation, what outperforms our previous thermal face solution based on CNN [26], where we assumed the strictly defined head pose during data acquisition. In addition, we proved that the examined model can be adapted to thermal imagery and achieve a high classification accuracy (99.92%). By providing single network for face classification in different scenarios, we eliminated the need for additional compensation techniques using landmarks or position annotations, what makes the proposed approach suitable for resource-constrained smart home devices. In future work, we want to perform more tests in a variety of experimental conditions to collect more diverse data. Also, we want to train models with augmented data including image rotations, to examine if it helps to improve CNN accuracy.

References

[1] Kamiyama S., "The Super Aged Society Japan's Kaiteki Institute Studies How to Keep the Elderly Healthy and Active", IEEE EMBS Pulse, 03/04 2014.

[2] Villa, A., Bellomo, D. "Performance Evaluation of Local Healthcare Systems by Applying Industrial Management Methods" In Health Care Management (WHCM), IEEE Workshop 2010, pp. 1-5.

[3] Mubarakov A., Zhengis Y., Kho Y. H., "Assistive Healthcare Home Monitoring System for Elderly People" IEEE 10th International Conference on Application of Information and Communication Technologies (AICT), Baku, 2016, pp. 1-5.

[4] Bujnowski A., Ruminski J., Przystup P., Czuszynski K., Kocejko T., "Self-Diagnostics Using Smart Glasses - Preliminary Study" 9th International Conference on Human System Interactions (HSI), Portsmouth, 2016, pp. 511-517.

[5] Lee S.J., Kim M. H., "KoMIPS: A Web-Based Medical Image Processing System for Telemedicine Applications" TENCON '02. Proc. IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, 2002, pp. 569-572 vol.1.

[6] Ruminski J., Kwasniewska A., "Evaluation of Respiration Rate Using Thermal Imaging in Mobile Conditions" Application of Infrared to Biomedical Sciences, pp. 311-346. Springer Singapore 2017.

[7] Etehadtavakol M., Ng E. Y., 2017. "Potential of Thermography in Pain Diagnosing and Treatment Monitoring" In Application of Infrared to Biomedical Sciences, pp. 19-32. Springer, Singapore.

[8] Thiruvengadam J, Anburajan M, Menaka M, Venkatraman B. "Potential of Thermal Imaging as a Tool for Prediction of Cardiovascular Disease", Journal of Medical Physics/Association of Medical Physicists of India. 2014 Apr;39(2):98.

[9] Lee Y, Paeng S, Farhadi H, Lee W, Kim S, Lee K, "The Effectiveness of Infrared Thermography in Patients with Whiplash Injury" Journal of Korean Neurosurgical Society, 57(4), pp.283-288, 2015.

[10] Prendergast P.M. "Anatomy of the Face and Neck" Shiffman M., Di Giuseppe A.(eds) Cosmetic Surgery Springer, Berlin Heidelberg, 2013

[11] Dong B., Biswas S., "Swallow Monitoring Through Apnea Detection in Breathing Signal" Annual International Conference of the IEEE EMBS, San Diego, CA, 2012, pp. 6341-6344.

[12] Cho, Y., Bianchi-Berthouze, N., Julier, S.J., "DeepBreath: Deep Learning of Breathing Patterns for Automatic Stress Recognition using Low-Cost Thermal Imaging in Unconstrained Settings", arXiv preprint arXiv:1708.06026.

[13] Lewandowska, M., Rumiński, J., Kocejko, T. Nowak, J., "Measuring Pulse Rate with A Webcam a Non-Contact Method for Evaluating Cardiac Activity" In Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on pp. 405-410. IEEE.

[14] Bellantonio M., Haque M., Rodriguez P., Nasrollahi K., Telve T., Escalera S., Gonzalez J., Moeslund T., Rasti P., Anbarjafari G., "Spatio-temporal Pain Recognition in CNN-Based Super-Resolved Facial Images" in: Nasrollahi K. et al. (eds) Video Analytics. Face and Facial Expression Recognition and Audience Measurement. FFER 2016, VAAM 2016. Lecture Notes in Computer Science, vol. 10165. Springer, Cham, 2017.

[15] He S., Soraghan J.J., O'Reilly B.F., "Objective Grading of Facial Paralysis Using Local Binary Patterns in Video Processing" in: 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, BC, pp. 4805-4808.

[16] Viola, P., Jones, M., "Rapid Object Detection Using a Boosted Cascade of Simple Features." In Computer Vision and Pattern Recognition, CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on vol. 1, pp. I-I. IEEE.

[17] Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G., "A Convolutional Neural Network Cascade for Face Detection" In Proc.of the IEEE Conference on Computer Vision and Pattern Recognition pp. 5325-5334

[18] Farfade, S.S., Saberian, M.J., Li, L.J., "Multi-View Face Detection Using Deep Convolutional Neural Networks", In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval pp. 643-650, ACM, 2015.

[19] Sun, Y., Wang, X., Tang, X., "Deep Convolutional Network Cascade for Facial Point Detection" in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on pp. 3476-3483.

[20] Cheng G., Zhou P., Han J., "Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images" in IEEE Transactions on Geoscience and Remote Sensing, Dec. 2016, vol. 54, no. 12, pp. 7405-7415.

[21] Breitenstein M. D., Kuettel D., Weise T., Van Gool L., Pfister H. "Real-Time Face Pose Estimation from Single Range Images" Computer Vision and Pattern Recognition, CVPR 2008. IEEE Conference on. IEEE, 2008.

[22] Ghiass, R. S., Arandjelović, O., Laurendeau, D., "Highly Accurate and Fully Automatic Head Pose Estimation from A Low-Quality Consumer-Level RGB-D Sensor" in Proceedings of the 2nd Workshop on Computational Models of Social Interactions: Human-Computer-Media Communication, 2015, pp. 25-34

[23] McCall R., Louveton N., Rumiński J., D2.1 The Specification and Overall Requirements of the eGlasses Platform. Technical Report, Univ. of Luxembourg, [http://orbilu.uni.lu/handle/10993/16763], (ISBN: 978-2-87971-125-6), Accessed Nov. 25, 2017

[24] Sabour S., Frosst N., Hinton G. E., "Dynamic routing between capsules" Advances in Neural Information Processing Systems, 2017.

[25] Xi, E., Bing, S., Jin, Y., "Capsule Network Performance on Complex Data" arXiv preprint arXiv:1712.03480, 2017.

[26] Kwasniewska A., Ruminski J., Czuszynski K., Szankin M., "Real-time Facial Features Detection from Low Resolution Thermal Images with Deep Classification Models", Journal of Medical Imaging and Health Informatics 2018, in print

[27] Szegedy, C., et al., "Rethinking the Inception Architecture for Computer Vision", CoRR abs/1512.00567 2015.